GUM-SAGE: A Novel Dataset and Approach for Graded Entity Salience Prediction

Jessica Lin and **Amir Zeldes**
Department of Linguistics
Georgetown University
{yl1290, amir.zeldes}@georgetown.edu

Abstract

Determining and ranking the most salient entities in a text is critical for user-facing systems, especially as users increasingly rely on models to interpret long documents they only partially read. Graded entity salience addresses this need by assigning entities scores that reflect their relative importance in a text. Existing approaches fall into two main categories: subjective judgments of salience, which allow for gradient scoring but lack consistency, and summarization-based methods, which define salience as mention-worthiness in a summary, promoting explainability but limiting outputs to binary labels (entities are either summary-worthy or not). In this paper, we introduce a novel approach for graded entity salience that combines the strengths of both approaches. Using an English dataset spanning 12 spoken and written genres, we collect 5 summaries per document and calculate each entity's salience score based on its presence across these summaries. Our approach shows stronger correlation with scores based on human summaries and alignments, and outperforms existing techniques, including LLMs. We release our data and code at https://github.com/jl908069/gum_sum_salience[1] to support further research on graded salient entity extraction.

1 Introduction

Salient entity extraction (SEE) is the task of identifying the most central entities mentioned in an arbitrary document in a given language, based on their contribution to the overall meaning of the document (Gamon et al., 2013). SEE has a range of applications, for example in news search and analysis, as well as summarization (Asgarieh et al., 2024), since users may want to categorize articles based on mentioned salient (but not non-salient) entities, or ensure that summary information focuses on salient, rather than tangential entities.

Two key challenges inherent in work on SEE are the gradualness and subjectivity involved in human salience judgments. For example, in a biography of Albert Einstein, Einstein is likely to be the most salient person, but other people mentioned can still be more salient, e.g. Danish physicist Niels Bohr, with whom Einstein entered into prominent debates, or less so, e.g. Einstein's uncle Jakob – both are mentioned in Einstein's Wikipedia article, the latter only briefly but the former 11 times. Although human raters are likely to easily agree that Bohr is more salient in Einstein's Wikipedia page than Jakob Einstein, there are many subtle cases on which they disagree: Dojchinovski et al. (2016) showed that crowdsourced entity salience annotations contained nearly 20% of labels that had to be discarded as 'untrustworthy', and achieved only around 63% agreement.

Other approaches opt to exploit additional properties of documents for a more operationalizable definition of salience, for example based on hyperlinks in the document (Wu et al., 2020) or the mention of entities in a summary of the document (Dunietz and Gillick, 2014; Lin and Zeldes, 2024). While these approaches reach much higher agreement (over 97% in Lin and Zeldes 2024), they are limited to single, binary judgments (an entity either appears in a summary or not, is either hyperlinked or not, etc.) and are more sensitive to variability in the underlying properties (many documents contain no links, a link may or may not be added, the summary could have been different, etc.).

In this paper, we aim to combine the benefits of clearly operationalized approaches to entity salience, specifically in the paradigm of summary-based salience, with the advantages of graded salience labels derived from multiple aggregated sources. Our approach relies on collecting mul-

[1]Data is also available from https://github.com/amir-zeldes/gum.

Findings of the Association for Computational Linguistics: ACL 2025, pages 438–455
July 27 - August 1, 2025

tiple, separate summaries of each document, and aligning mentions to the original text to create numerical salience scores based on the number of summaries mentioning each entity (e.g. 5/5 summaries would mention Einstein in his biography, some might mention Bohr, and probably none would mention his uncle). To the best of our knowledge, this is the first attempt to cast summarization-based salience as a regression, rather than a classification problem.

The main contributions of this paper are:

- A novel approach to graded summary-based entity salience prediction, demonstrating a nearly 17-point improvement in F1 score and better correlation with human summary and alignment-based salience scores compared to leading LLMs
- A new dataset based on the openly available UD English GUM corpus (Zeldes, 2017), semi-automatically enriched with 5 aligned summaries and graded salience scores for all named and non-named entities
- A thorough evaluation of models and systems used to create our data, including fully manually constructed test and dev sets
- Analysis of error patterns in models' entity salience predictions and their correlation with human annotated salience

2 Related Work

The increasing importance of SEE is reflected in the expanding number of annotated datasets, employing different strategies for entity recognition and salience labeling (see Table 1). However, building a reliable dataset with consistent entity salience annotations remains a significant challenge for a number of reasons, including lack of reliable and exhaustive entity annotations, the absence of consistent guidelines for entity salience, subjectivity in the assignment of salience scores, and limited data availability across text genres, with previous work focusing almost only on news and Wikipedia material.

To ensure the reliability of an entity salience dataset, the first step is to adopt a robust method for identifying entities within a document. Some work (Dunietz and Gillick, 2014; Gamon et al., 2013) has utilized multi-step automatic pipelines (including NP extraction, coreference resolution, and a named entity recognizer) to identify entities, while others (Dojchinovski et al., 2016; Trani et al., 2018; Wu et al., 2020) have undertaken manual annotation. The latter is potentially more accurate but expensive, while NLP pipelines are cheap but may propagate errors to later steps. Furthermore, salient entities in a document are not necessarily named entities, but also non-named ones. Most previous datasets (Bullough et al., 2024; Dojchinovski et al., 2016; Dunietz and Gillick, 2014; Gamon et al., 2013; Maddela et al., 2022; Sekulic et al., 2024; Trani et al., 2018; Wu et al., 2020) include only named entities, leaving out common noun entities that may be salient to humans – in fact, some documents contain no named entities, but we would still assume some of the non-named entities will be more salient than others.

The second challenge in building a reliable entity salience dataset is how to minimize noise in salience labels and apply consistent guidelines. Entity salience labels have been derived either through crowdsourcing, gathering ratings from multiple non-expert raters to determine salient entities (Bullough et al., 2024; Dojchinovski et al., 2016; Maddela et al., 2022; Sekulic et al., 2024; Trani et al., 2018), or by employing proxy methods such as abstracts or writer-assigned Wikinews categories (Gamon et al., 2013; Dunietz and Gillick, 2014; Wu et al., 2020). While crowdsourcing can surpass automated methods in performance, it is inherently noisy and prone to bias, as opinions on what is salient can be unpredictable (Maddela et al., 2022). On the other hand, using proxies has been shown to be less noisy (i.e. more reproducible), but can suffer from a lack of reliability in the proxies themselves. For example, the NYT (**N**ew **Y**ork **T**imes)-salience dataset (Dunietz and Gillick, 2014) relied on found news abstracts to identify salient entities. This approach has several limitations: First, the derived salience labels may be less reliable due to the lack of clear and consistent guidelines for summaries (e.g. length, style). Second, relying on news article abstracts restricts the dataset to certain types of genres (i.e. news).

In this study, we adopt a regimented approach similar to the NYT salience corpus (Dunietz and Gillick 2014), which identifies salient entities using summaries. To improve annotation reliability, we crowdsourced summaries across 12 English genres, following very specific guidelines

Datasets	Multi-Genre	Multi-types	# of Documents	# of Entities	% of Salient Entities	Entity Annotations	Salience Labels	Graded Salience
MDA (Gamon et al., 2013)	✓	✗	≈ 50,000	2,414	< 5 %	proprietary NLP pipeline	soft labeling	✓
NYT-Salience (Dunietz and Gillick, 2014)	✗	✗	110,639	2,229,728	≈ 14 %	proprietary NLP pipeline	automated	✗
Reuters-128 Salience (Dojchinovski et al., 2016)	✗	✗	128	4,429	18%	manual annotation	crowdsourcing	✓
The Wikinews dataset (Trani et al., 2018)	✗	✗	365	≈ 4,400	≈ 10 %	manual annotation	crowdsourcing	✓
WN-Salience (Wu et al., 2020)	✗	✗	6,968	888	16%	manual annotation	automatic derivation	✗
EntSUM (Maddela et al., 2022)	✗	✗	693	7,854	39%	semi-automated	crowdsourcing	✓
WikiQA-Salience (Bullough et al., 2024)	✗	✗	687 Q/A pairs	2,113	≈ 52% high salience	open source NLP library	crowdsourcing	✓
CIS Entity Salience (Sekulic et al., 2024)	✗	✗	120 Q/A pairs	∼400	≈ 63%	open source NLP library	crowdsourcing	✓
GUMsley (Lin and Zeldes, 2024)	✓	✓	213	29,899	7%	manual annotation	semi-automated	✗

Table 1: Statistics of existing entity salience datasets. The column `Multi-types` shows whether the dataset covers diverse types of entities (named entities, non-named entities, wiki-linked entities) and NPs (e.g., verbal NPs).

proposed by Liu and Zeldes (2023a). This method increases the reliability of the summaries as proxies for salience, and in turn the consistency of salient entity annotations, compared to reliance on found abstracts limited to news articles, which were not designed for this task. Additionally, it allows us to obtain graded salience judgments without sacrificing this reliability.

3 Dataset

Our dataset, called GUM-SAGE (**GUM**-based **S**ummary **A**ligned **G**raded **E**ntities) is based on the GUM corpus (Zeldes, 2017), an open-access manually annotated, multilayer resource for English. The corpus spans over 200K tokens across 12 different text genres (see Table 2), and includes Universal Dependencies (UD) parses (de Marneffe et al., 2021), detailed entity annotations such as entity types and Wikification links (Lin and Zeldes, 2021), coreference resolution (Zhu et al., 2021), and discourse parses (Liu and Zeldes, 2023b). Additionally, the dataset provides an expert-written summary for each document (Liu and Zeldes 2023a), aligned to the entity annotations for entities mentioned in the summary (Lin and Zeldes, 2024), which we leverage for evaluation below.

In this paper we add entity salience scores (0-5) for all named and non-named entities in the data using an SEE pipeline with two components: Summary Crowdsourcing & Generation (Section 3.1) and Entity Alignment (Section 3.2). We assume that it is difficult to summarize a text without mentioning its most salient entities, and that salient entities will therefore tend to appear in summaries, while spuriously mentioned entities will not recur in many summaries. The SEE pipeline therefore collects multiple summaries per document, aligning mentions to assign salience scores based on the number of summaries mentioning the entity. We evaluate the accuracy of our approach in Section 4.

3.1 Summary Crowdsourcing & Generation

Each document in GUM is already accompanied by a single expert-written summary, and an additional second human-written summary is provided for each of the 24 test documents (Liu and Zeldes, 2023a). However because our approach to salience is based heavily on summary content, which can vary, a single summary may be inadequate to identify salient entities within a document, by either missing some salient entities, or containing spurious ones. We therefore crowdsource or generate summaries for our data.

Summary Crowdsourcing In the summary crowdsourcing task, each annotator is asked to read eight documents from different genres before writing a one-sentence summary for the document, which should 'substitute reading the text', focus on 'who did what to whom', and, space allowing, 'when, where and how', but may not exceed 380 characters, following Liu and Zeldes (2023a). Annotators were also instructed not to mention anything not mentioned in the document, and to adhere as closely as possible to the document's vocabulary and phrasing. All of the crowdsourced summaries were manually checked by one of the authors to ensure that they follow the guidelines. Most of the summaries did so, though a small portion deviated in two key areas: (i) Mentioning facts not mentioned in the text, such as speaker names not explicitly stated but identifiable from context; (ii) Using shell nouns like "this reddit post", which should generally be avoided if they are not unambiguously identifiable in the text (e.g. a writer states this is a reddit post). Any summaries that did not comply with guidelines were minimally manually corrected to maintain consistency, without otherwise altering their meaning.

	academic	bio	conversation	fiction	interview	news	reddit	speech	textbook	vlog	voyage	wikihow	TOTAL
Documents	18	20	14	19	19	23	18	15	15	15	18	19	**213**
Tokens	17,905	18,554	14,307	18,003	16,504	21,767	17,986	13,195	14,451	14,784	17,984	18,341	**203,781**
Mentions	5,045	5,768	4,094	4,974	5,211	4,720	4,544	4,847	4,719	4,499	4,471	4,468	**57,360**
Entities	3,251	3,324	1,363	2,352	2,642	2,579	2,364	2,573	2,885	1,626	2,957	2,384	**32,300**
Avg # of Entities	181	166	97	124	139	112	131	172	192	108	164	125	**148**
% Salient Entities	6.3	9.1	9.4	8.0	9.6	11.6	12.2	14.4	15.8	16.8	19.4	32.9	**13.8**
% of Top1 Salient Entities	0.9	1.2	3.2	1.5	2.0	2.3	1.5	2.1	1.8	2.4	2.5	3.6	**2.1**
% of Top3 Salient Entities	1.9	2.8	6.6	3.5	3.7	4.6	4.7	4.6	4.9	6.5	5.1	9.9	**4.9**

Table 2: Overview of GUM-SAGE. Top1 salient entities are those with a score of 5; Top3 refers to entities with scores of 3, 4, or 5. % salient entities = number of all salient entities (score 1-5) / total number of entities. Avg entities per summary = # of entities / # of documents in the genre.

Summary Generation We select four recent LLMs – `GPT-4o` (`OpenAI, 2024`), `Claude 3.5 Sonnet` (Anthropic, 2024), `Llama 3.2 3B Instruct` (Meta AI, 2024), and `Qwen 2.5 7B Instruct` (Qwen Team, 2024) – to create four "silver" summaries[2] for each of the 165 training set documents, matching the length and style of GUM summaries. Along with one gold summary per document, this resulted in five summaries per document. All models were instructed to produce a one-sentence summary and were given examples from the dev set for the genre in question[3].

Summaries violating the length limit were resolved by re-prompting the LLM (in the same session) to abbreviate to the required length. Finally, minimal automatic corrections were applied, such as replacing periods with semicolons if models outputted more than one sentence, in order to ensure compatibility with the manual gold dev and test sets of the GUM-SAGE.

3.2 Entity Alignment

We aligned mentions in human-written and system-generated summaries with those in the document using several methods, from rule-based methods to NLP pipelines, to prompt-based LLMs, as well as manual alignment for the dev/test data[4].

- **String Match**: To achieve high precision in aligning mentions to summaries with corresponding mentions, we use exact and partial string matching, iterating over the gold entity annotations in each GUM document: for multi-word mentions (>2 tokens), we allow for partial matching when more than 3 contained tokens appear exactly in the summary, excluding stop words. For example, if a summary mentions 'the prevalence of racial discrimination in the United States', and the document mentions 'The prevalence of discrimination across racial groups in contemporary America', the module considers the entity to appear in the summary due to a substring match (*prevalence, racial, discrimination*). This approach provides flexibility in cases where the phrasing of mentions between the summary and the document may differ.

- **Stanza Coreference Model** (Liu et al., 2024): We concatenate each summary to its document and use the Stanza coreference model, trained on CorefUD (Nedoluzhko et al., 2022) with `XLM-RoBERTa-large` (Conneau et al., 2020), to align mentions between documents and their summaries. Each mention in the document predicted to be in a coreference cluster with any mention in the summary is considered to be included in the summary, and therefore salient.

- **GPT-4o** (OpenAI, 2024): We used[5] GPT-4o to determine whether each entity in the document is also mentioned in the corresponding summary, prompting it to take synonyms and alternative phrases into account. We expect GPT-4o to achieve high precision, given its capability to perform at human-level accuracy on various benchmarks. However in initial experiments, feeding the model all document entities at once harmed recall, with the model struggling to identify all relevant matches in a large batch. To optimize recall, we feed the model batches of 15-20 document entities at a time, asking it to match these with those in the summary. This

[2] See Section C for details on the quantitative evaluation of generated summaries.

[3] See Appendix B.1 for prompt details.

[4] See Appendix D for details on the interface we used.

[5] See Appendix E for prompt details.

process is repeated through multiple non-overlapping queries for each document, ensuring all entities are evaluated without overwhelming the model. Since we elicit binary judgments per entity, the model's responses can easily be parsed to extract an alignment, which we aggregate across queries.

- **Ensemble Learning**: To enhance alignment accuracy, we train a logistic regression model on the manually corrected dev set to predict entity salience. The model uses binary predictions from the String Match module, Stanza, and GPT4o as features, and includes other features available from GUM annotations, such as entity type (person, organization, etc.), genre, and document position, achieving stronger performance than any individual module (Table 3). The position feature is defined as the entity's order in the document, e.g. the first entity has a position of 1, etc.

4 Evaluating Alignment

Table 3 presents alignment scores for different alignment methods in Section 3.2 in matching entities shared between documents and summaries. Note that reported scores refer to the test partition only, since we use the dev partition as training data for the ensemble.[6]

Alignment component	Micro Average			Macro Average		
	P	R	F	P	R	F
Flan-T5-XL	0.95	0.08	0.14	0.57	0.09	0.15
String match	0.98	0.39	0.56	0.97	0.37	0.52
GPT-4o (avg 3 runs)	0.80	0.71	0.75	0.82	0.72	0.76
Stanza Coref	0.73	0.82	0.77	0.74	0.82	0.77
Ensemble learning	0.98	0.98	**0.98**	0.98	0.92	**0.95**
Ensemble learning (for the positive class)	0.98 / 0.83 / 0.90					

Table 3: Alignment scores for different alignment components on the test set of GUM-SAGE. The best performing F1 scores are in **bold**.

Interestingly, the Stanza coreference model achieves marginally better overall performance compared to GPT-4o, particularly in *conversation* and *vlog* genres (Figure 1), despite GPT-4o's vastly larger training data and parameter count,

[6]The alignment for both the test and dev partitions was manually corrected for all summaries using the interface described in Appendix D.

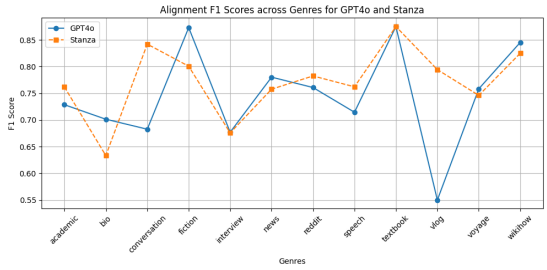


Figure 1: Alignment F1 Scores on the test set across genres for GPT4o and Stanza.

and both models having access to the same inputs (the document, including gold speaker labels for spoken data). We speculate that Stanza's dedicated coreference architecture benefits from optimization for mention detection, a critical task where prompt-based LLMs like GPT-4o struggle, as noted by Sundar et al. (2024) and Le and Ritter (2023), perhaps due to the less natural task of exhaustively identifying all mentions, including nested ones. Conversational genres like unscripted conversations and vlogs, which involve frequent pronoun use, informal language, and rapid speaker shifts seem to be particularly challenging. Secondly, Stanza benefits from its training on CorefUD (Nedoluzhko et al., 2022), which includes GUM as the largest open English coreference dataset, making the model familiar with GUM's coreference span definitions (see Zeldes 2022), and allowing it to excel in the genres included in the corpus (however note that the Stanza model could not have been exposed to our summaries, which are novel, and is not trained on the test data).

Although no single module achieves satisfactory performance for creating a new benchmark dataset — a key goal of this paper — the ensemble learning approach in Table 3 stands out by leveraging the distinct strengths of each component method. While string matching and GPT-4o achieve high precision by identifying clear entity matches, Stanza's coreference architecture provides higher recall by capturing more paraphrases. The ensemble approach combines these complementary signals - precise explicit matches and broader referential coverage - alongside linguistic features to make more robust predictions about entity salience. Although the final micro F1 score of 0.98 seems very high, we note that this is in part due to the frequency of negative judgments

(most document mentions do not appear in summaries), and the positive class F1 is lower, at 0.9. We also note that the task is easier than might be expected, since our guidelines explicitly ask summary writers to adhere to things mentioned in the documents, and to keep phrasing similar to the source material.

5 Predicting salient entities

The purpose of the dataset created in the previous sections is to train and evaluate systems on the task of gradient entity salience prediction. To evaluate contemporary models on this task, we prompt[7] LLMs to extract (i) salient entities from a document and (ii) a salience score from 1 to 5 for each predicted salient entity (with entities absent from the prediction corresponding to a score of 0).

5.1 Models

Zero-shot/Few-shot LLMs We used GPT-4o (OpenAI, 2024) in both zero-shot and few-shot settings to identify salient entities within a document and assign a salience score between 1 and 5 to each predicted entity. For the few-shot setting, we provide 3 randomly selected documents from the dev set, along with their salience scores, as in-context examples to guide the model's predictions. Additionally, we evaluated other instruction-based models, including `Llama-3.2-3B-Instruct` (Meta AI, 2024) and `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), which are tuned for tasks like information retrieval and text summarization. These models are prompted to extract salient entities, such as people and organizations, and assign salience scores from 1 to 5. We chose instruction-based models over base models due to their fine-tuning on retrieval tasks, which aligns with our goal of retrieving and ranking salient entities.

Stanza & Ensemble Approach We evaluated two approaches: the Stanza model and our proposed ensemble method. Following our definition of salient entities as those present in both documents and summaries, we extracted binary salience labels from each summary and aggregated them across all five model-generated summaries[8] to derive a 1-5 salience score per entity.

Position Baseline We built our baseline model on the simple and naive assumption that entities mentioned earlier in a document tend to be more salient, as important information is often introduced early (Dojchinovski et al., 2016; Liu et al., 2018). Entities are assigned salience scores based on their first appearance within the document. The document is divided into sections, where entities in the first 10% of sentences are scored 5, with scores decreasing by 1 for each subsequent 10-20% segment, down to 0 for the last 20%. This method provides a simple, interpretable baseline that reflects the salience of entities based on their position in the document.

5.2 Evaluation Metrics

In our evaluation experiment, we evaluate the correlation between predicted and human-aggregated salience scores using two metrics: Spearman's ρ and Root Mean Square Error (RMSE), which capture different aspects of prediction quality. Spearman's ρ, ranging from -1 to +1, measures the strength and direction of the monotonic relationship between the rankings of predicted and human scores. For example, if two entities are ranked similarly by both the model and the annotators (even if the raw scores differ), we will observe a high Spearman correlation. In contrast, RMSE measures the absolute difference between predicted and annotated scores, providing insight into how far off the predictions are on average, regardless of the ranking; if a model consistently predicts scores that are close to the annotated values, RMSE will be low.

To evaluate the effectiveness of the models in identifying entities with high salience, we compute precision, recall, and F1 scores for the top1 (entities with a score of 5 out of 5) and top3 (entities with a score of 3, 4, 5 out of 5) entity ranks to capture the model's ability in detecting the most salient entities. We expect a positive Spearman ρ, indicating a correlation in ranking performance. For precision, recall, and F1 scores, we expect the highest scores for top1 entities, with the scores decreasing as we move to top3. This is because we expect models to be more accurate at identifying highly salient entities (score 5), and to degrade as the salience level becomes less distinctive.

[7] Settings and prompt in Appendix E

[8] For fairness, we used 5 model-generated summaries for evaluation (Section 6), using the same models and methods from Section 3.1, despite having human-written ones available. This allows us to evaluate how well our approaches handle end-to-end salience prediction from documents alone, ensuring a fair comparison with LLMs using the same inputs.

6 Evaluation

In this section, we evaluate model performance on the graded entity salience prediction task end to end, measuring the correlation between predicted and human-aggregated salience scores using only the document as input (Section 6.1). In Section 6.2, we focus on error analysis using GPT4o predictions; analyses and predictions of other models are included in Appendix F.

6.1 Model Performance

The results in Table 4 show that the Ensemble approach achieves the highest Spearman's ρ (0.54), indicating the strongest alignment with human salience rankings, while Stanza achieves the lowest RMSE (1.03). Both methods outperform GPT-4o in zero-shot (ρ: 0.229, RMSE: 1.143) and three-shot settings (0.254, 1.111), suggesting that despite their generalization capabilities, LLMs may struggle with fine-grained salience prediction tasks compared to task-specific methods. Additionally, Wilcoxon signed-rank tests confirm that the improvements in rank correlation over GPT-4o (3-shot) are statistically significant for both Stanza and Ensemble ($p < 0.001$), supporting the robustness of our approach. While these results are promising, with Ensemble showing substantial alignment with rankings based on human-generated summaries and alignments, and Stanza's predictions deviating by approximately one point on the 0 to 5 scale, there remains room for improvement on this challenging task.

Table 5 evaluates the model's performance in identifying salient entities under two evaluation settings: Top1 (only entities with a salience score of 5) and Top3 (entities with a salience score of 3, 4, or 5). Importantly, these evaluations are restricted to salient entities only and do not include non-salient entities (salience score = 0) as negatives. As such, the reported scores reflect the model's ability to prioritize the most salient entities without being affected by the large number of non-salient mentions in the dataset.

Overall, the Top1 setting emphasizes strict precision on highly salient entities, whereas Top3 favors a more recall-oriented evaluation. Precision is higher under the Top1 setting, where only the most salient entities are considered, while recall increases under the more inclusive Top3 setting. This trade-off arises because the Top3 evaluation rewards models for capturing additional, moderately salient entities (scores 3 or 4), but at the cost of introducing more false positives, thereby reducing precision.

Our Stanza and Ensemble approaches excel in recall for Top1 salient entities, with the Ensemble achieving the highest recall (0.755). This is because our methods are more lenient, aiming to capture as many salient entities as possible, which increases recall but also introduces false positives, thereby lowering precision. In contrast, LLMs like GPT4o are more conservative in their predictions, focusing on a narrower set of entities that are clearly salient. This conservatism results in higher precision (e.g., 0.548 for GPT4o few-shot) but inherently limits their recall, as they may overlook less obvious yet salient entities.

For the Top3 scenario, our Ensemble approach outperforms all models, achieving the highest recall (0.61) and F1 score (0.527), effectively balancing precision and recall to capture moderately salient entities. The Stanza method also performs well, achieving the second-highest precision (0.424) among all models and a recall (0.474) that outperforms several LLMs. In contrast, LLMs like GPT4o underperform, with both precision and recall falling short of the Ensemble, highlighting the advantage of our task-specific strategies in identifying a wider range of salient entities.

Although our Ensemble approach achieves the highest performance in the Top3 scenario, with an F1 score of 0.52 and a spearman correlation of 0.54, these results highlight the fundamental challenge of consistently identifying and ranking entities by their salience in text.

6.2 Error Analysis

The confusion matrix in Figure 2 shows that GPT4o is more confident in predicting highly salient entities but struggles to differentiate less salient ones, often overpredicting moderate salience scores (3 and 4) for entities that are actually less salient (1 and 2). This pattern is observed in other models as well (Figure 18). The model performs best with highly salient entities (score 5) but tends to misclassify low-salience entities (score 1) into higher salience categories like 3 or 4, likely because moderate scores are a safer prediction for uncertain cases, reducing extreme errors but introducing more moderate misclassifications. We speculate that models' natural language understanding does not calibrate them well

	Spearman ρ (95% CI)	RMSE (95% CI)	Wilcoxon Test (Spearman ρ vs GPT4o 3-shot)
Position Baseline	0.153 (0.103, 0.208)	2.554 (2.471, 2.636)	Spearman: * (p<0.05)
GPT4o 3 shot	0.254 (0.208, 0.300)	1.111 (1.044, 1.182)	-
GPT4o zero shot	0.229 (0.179, 0.280)	1.143 (1.075, 1.213)	ns
Llama-3.2-3B-Instruct	0.223 (0.167, 0.281)	1.296 (1.211, 1.397)	ns
Mistral-7B-Instruct-v0.3	0.254 (0.202, 0.307)	1.206 (1.124, 1.286)	ns
Stanza (our approach)	0.384 (0.324, 0.437)	**1.031** (0.932, 1.131)	Spearman: *** (p<0.001)
Ensemble (our approach)	**0.540** (0.486, 0.589)	1.067 (0.933, 1.208)	Spearman: *** (p<0.001)

Table 4: Performance of all models on the graded entity salience prediction task (test set). We report Spearman's ρ and RMSE, each with 95% confidence intervals. Our two approaches—Stanza and Ensemble—significantly outperform GPT-4o (3-shot) in rank correlation (Spearman's ρ), as confirmed by a Wilcoxon signed-rank test ($p < 0.001$). The Ensemble model achieves the strongest overall performance. Best scores per column are shown in **bold**. "ns" indicates a non-significant difference at $p \geq 0.05$.

Model	P@top1	R@top1	F@top1
GPT4o 3 shot	**0.548**	0.321	0.405
GPT4o	0.541	0.377	**0.444**
Mistral 7b	0.415	0.321	0.362
Llama 3-2	0.302	0.359	0.328
Stanza	0.239	0.491	0.321
Ensemble	0.242	**0.755**	0.367
Position Baseline	0.031	0.208	0.054
Avg	0.331	0.405	0.326
Model	**P@top3**	**R@top3**	**F@top3**
GPT4o 3 shot	0.278	0.513	0.361
GPT4o	0.255	0.481	0.333
Mistral 7b	0.205	0.455	0.283
Llama 3-2	0.183	0.442	0.259
Stanza	0.424	0.474	0.448
Ensemble	**0.463**	**0.610**	**0.527**
Position Baseline	0.031	0.305	0.057
Avg	0.263	0.469	0.324

Table 5: Precision, recall, and F1 scores for all models on the test set. @Top1 means only the entities with a score of 5 are considered; @Top3 means the entities with a score of 3 or 4 or 5 are considered. The highest scores in each scenario are in **bold**.

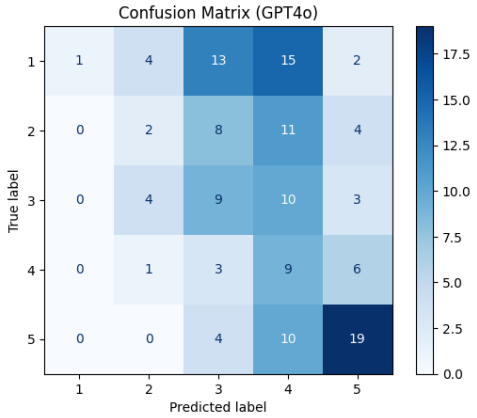


Figure 2: Confusion matrix for GPT4o

to the scale implied by the data.

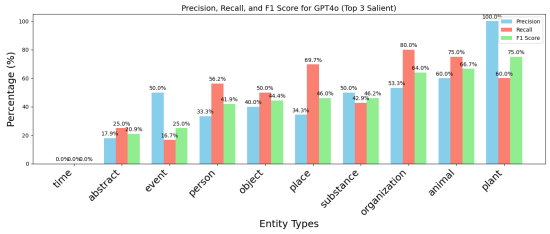


Figure 3: Performance of GPT4o across entity types in identifying the top 3 salient entities. Precision (the blue bar) measures the percentage of predicted salient entities for a given type that are correct, recall (the red bar) measures the percentage of actual salient entities that the model successfully predicts.

The results in Figure 3 shows GPT4o's varying performance across entity types when predicting the Top 3 salient entities. The model performs well on concrete and named entities like ANIMAL, PLANT and ORGANIZATION, which exhibit higher precision, recall, and F1 scores. Conversely, the model struggles with conceptual or contextual entities such as ABSTRACT, EVENT, and TIME, reflecting a need for deeper semantic understanding to predict salience effectively. For entities like PERSON and PLACE, the much higher recall compared to precision suggests overpredictions for these frequently occurring and easily identifiable entity types. This performance disparity[9] highlights the model's strengths with concrete entity types but its limitations in processing context-dependent or abstract entities.

The results in Figure 4 show that for GPT4o models, false positive (FP) and false negative (FN) counts decrease from the first to the second half of the document, with FP counts declining more steeply. This indicates that the model overpredicts

[9]Similar patterns can be seen in other models as well (Figure 19).

salience for entities mentioned early in the document, resulting in a higher FP rate in the first half. The steep FP decline suggests models are more conservative in the second half, likely due to fewer entities being introduced. The gentler FN decline reflects the smaller number of salient entities in the second half, giving models fewer opportunities to miss salient predictions.

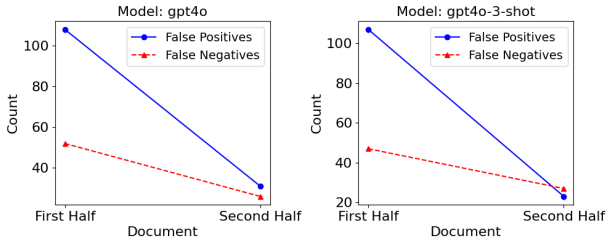


Figure 4: Interaction plot showing the counts of false positives (FP) and false negatives (FN) for GPT4o models across the first and second halves of documents.

7 Conclusion

In this paper, we introduced a novel approach to graded salient entity extraction, combining the strengths of human judgments and summary-based operationalization. By aligning multiple summaries for each document to their texts, we created a dataset with graded salience scores that balances gradient scoring with consistency and explainability. While predicting entity salience remains challenging even for powerful LLMs in zero-shot and few-shot settings, our novel SEE approach demonstrates promising results by outperforming these models. Our analysis shows that models perform well on concrete entities but poorly on abstract ones (abstract notions, time), with bias toward common entities and those appearing early in text.

The data set and approach presented here establish a foundation for improving salience modeling in summarization and information retrieval. We are also planning to add the same graded annotations described in this paper to upcoming editions of the GUM corpus currently covering 16 genres (with the addition of court transcripts, essays, letters and podcasts), as well as the out-of-domain challenge test set GENTLE (the Genre Tests for Linguistic Evaluation corpus, Aoyama et al. 2023), with eight additional genres. Future work using this data should continue to address model biases and promote stability across genres.

Limitations

Our work has several limitations. First, the graded summary-based approach relies on multiple high-quality summaries per document (either human- or model-generated), which introduces scalability challenges—particularly in low-resource settings. Collecting and validating multiple summaries is resource-intensive and naturally constrains dataset size. While this setup is appropriate for evaluation and for establishing the gradient-based salience framework, we anticipate that human-written summaries may not be required when deploying the method in practice. As shown in Table 4 and 5, our methods perform competitively using only model-generated summaries. Moreover, recent advances in APIs have made it easier to generate multiple summaries at scale, reducing the overhead associated with querying different LLMs.

Second, we acknowledge the potential for pretraining contamination in LLMs, as our evaluation documents come from public sources. Specifically, since our dataset uses the open-source GUM corpus, some documents might have been part of models like GPT-4's pretraining data. However, we believe this concern primarily affects the evaluation of summarization quality, not the alignment-based salience prediction task. The summaries used for alignment are independently generated (by humans or models), and the alignment process involves matching *novel* summary content to entities in the source text, which makes it unlikely that the exact summary-document pairs were observed during pretraining. Furthermore, as shown in Figure 11, the low Self-BLEU scores across genres indicate that the summaries in our dataset are lexically diverse and not overly repetitive, supporting their effectiveness as a robust input for SEE even in the presence of potential pretraining overlap.

Third, our dataset is currently limited to English, the highest-resource language in NLP. The performance of pretrained models on both summarization and entity salience tasks would likely be substantially lower for other languages. While we believe the fundamental approach of using multiple summaries to grade entity salience should generalize across languages, this remains to be empirically verified, particularly for languages with different discourse structures or conventions around entity reference.

Finally, although the dataset spans 12 genres, domain-specific patterns of entity salience may

still be underrepresented. The semi-automatic alignment process, while scalable and robust, continues to require human validation to ensure accuracy. This may limit applicability in genres or domains where reference patterns or discourse organization differ substantially from those represented in our corpus.

Acknowledgments

The summary crowdsourcing study was supported by a GSAS-GradGov Research Project Award (GRPA), which funds graduate student research and professional development at Georgetown University. We are grateful to the following participants for their valuable contributions and thoughtful feedback during the summary crowdsourcing process (listed alphabetically by last name): Caroline Coggan, Jessica Cusi, Dan DeGenaro, Caroline Gish, Aniya Harris, Abby Killam, Lauren Levine, Cindy Li, Robbie Li, Cindy Luo, Todd McKay, Sophie Migacz, Anna Prince, Emma Rafkin, Eliza Rice, Wesley Scivetti, Devika Tiwari, Shira Wein.

References

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf.

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.

Eliyar Asgarieh, Kapil Thadani, and Neil O'Hare. 2024. Scalable detection of salient entities in news articles.

Benjamin Bullough, Harrison Lundberg, Chen Hu, and Weihang Xiao. 2024. Predicting entity salience in extremely short documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 50–64, Miami, Florida, US. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Milan Dojchinovski, Dinesh Reddy, Tomáš Kliegr, Tomáš Vitvar, and Harald Sack. 2016. Crowdsourced corpus with entity salience annotations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3307–3311.

Jesse Dunietz and Dan Gillick. 2014. A new entity salience task with millions of training examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 205–209.

Michael Gamon, Tae Yano, Xinying Song, Johnson Apacible, and Patrick Pantel. 2013. Identifying salient entities in web pages. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2375–2380.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Nghia T Le and Alan Ritter. 2023. Are large language models robust coreference resolvers? *arXiv preprint arXiv:2305.14489*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2021. WikiGUM: Exhaustive entity linking for wikification in 12 genres. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 170–175, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2024. GUMsley: Evaluating entity salience in summarization for 12 English genres. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2575–2588, St. Julian's, Malta. Association for Computational Linguistics.

Houjun Liu, John Bauer, Karel D'Oosterlinck, Christopher Potts, and Christopher D. Manning. 2024. MSCAW-coref: Multilingual, singleton and

conjunction-aware word-level coreference resolution. In *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–40, Miami. Association for Computational Linguistics.

Janet Yang Liu and Amir Zeldes. 2023a. GUMSum: Multi-genre data and evaluation for English abstractive summarization. In *Findings of ACL 2023*, Toronto.

Yang Janet Liu and Amir Zeldes. 2023b. Why can't discourse parsing generalize? A thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhengzhong Liu, Chenyan Xiong, Teruko Mitamura, and Eduard Hovy. 2018. Automatic event salience identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1226–1236, Brussels, Belgium. Association for Computational Linguistics.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. Accessed: October 22, 2024.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

OpenAI. 2024. Gpt-4 Technical Report. https://openai.com/index/gpt-4-research/.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ivan Sekulic, Krisztian Balog, and Fabio Crestani. 2024. Towards self-contained answers: Entity-based answer rewriting in conversational search. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 209–218.

Kawshik Sundar, Shubham Toshniwal, Makarand Tapaswi, and Vineet Gandhi. 2024. Major entity identification: A generalizable alternative to coreference resolution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11679–11695.

Salvatore Trani, Claudio Lucchese, R. Perego, David E. Losada, Diego Ceccarelli, and Salvatore Orlando. 2018. SEL: A unified algorithm for salient entity linking. *Computational Intelligence*, 34:2 – 29.

Chuan Wu, Evangelos Kanoulas, Maarten de Rijke, and Wei Lu. 2020. Wn-salience: A corpus of news articles with entity salience annotations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2095–2102.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2022. Can we fix the scope for coreference? Problems and solutions for benchmarks beyond OntoNotes. *Dialogue & Discourse*, 13(1):41–62.

Shuo Zhang and Amir Zeldes. 2017. GitDOX: A linked version controlled online XML editor for manuscript transcription. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2017)*, pages 619–623.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

A API costs

A.1 Summary Generation

For summary generation, we produced four one-line summaries for each of the 165 documents in the training set, totaling 660 requests. Using GPT-4o, priced at $2.50 per 1M input tokens and $10.00 per 1M output tokens[10], the cost was $0.90. For Claude 3.5 Sonnet, priced at $3 per 1M input tokens and $15 per 1M output tokens[11], the cost was $3.34. In total, the cost for summary generation was $4.24. Generating summaries with

[10] https://openai.com/api/pricing/

[11] https://www.anthropic.com/pricing#anthropic-api

Llama 3.2 and Qwen models using the Huggingface `transformers`[12] library did not incur any additional costs, as these models are open-source and locally hosted.

A.2 Automatic Alignment

The alignment task involved matching entities across document-summary pairs, requiring 6,660 requests processed using GPT-4o, with a total cost of $40.14.

B Summary Generation & Crowdsourcing Details

B.1 Summary Generation

Figure 5 shows the prompt to generate a one-sentence summary with all LLMs (`GPT4o`, `Claude 3.5 Sonnet`, `Llama 3.2 3B`, `Qwen 2.5 7B`). A max token of 120 is used to make sure the generated summary is not too long. All the other hyperparameters are in their default settings.

B.1 Prompt for LLMs for Summary Generation

Summarize the following article in 1 sentence. Make sure your summary is one sentence long and may not exceed 380 characters.
Example of summary style: {example}
Document: {doc_text}
Summary:

Figure 5: Prompt for LLMs for Summary Generation.

B.2 Summary Crowdsourcing

We recruited 24 graduate students at Georgetown University who are native speakers of English to write summaries for the test and dev set of GUMSAGE. Each writer, paid $22/hr (based on the pay rate of the 2023 / 2024 academic year for graduate students at Georgetown University), wrote 8 summaries, resulting in 192 human-written summaries in total. Each student received a Google Form to write summaries for their assigned texts, which can be viewed from an interface titled `GUM Full Text Reader` (Figure 6 and 7). Figure 6 and 7 show the interface for viewing articles in GUM. Figure 8 and 9 show the instructions given to the writers.

[12] See `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct` for the Llama model and `https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct` for the Qwen model used in this paper.

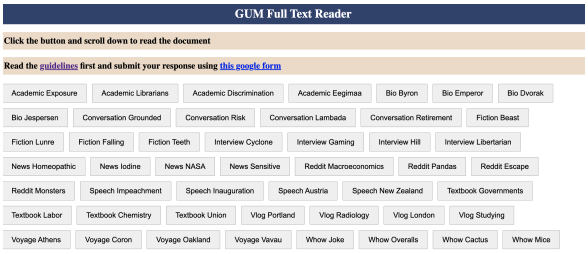


Figure 6: GUM Full Text Reader: The interface for viewing GUM articles.

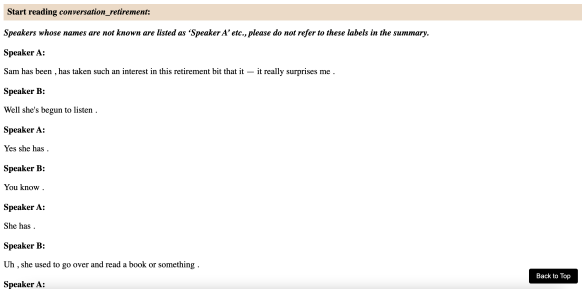


Figure 7: An example of GUM Full Text Reader.

C Automatic Evaluations of Summary Quality

To validate the quality of the summaries used for salient entity alignment in Section 3.2, we conducted automatic evaluations comparing the human-written and LLM-generated summaries across genres. ROUGE-1, ROUGE-2, and ROUGE-L scores (Lin, 2004) were computed using the human summary as reference (Figure 10). We found that LLM-generated summaries, particularly those from GPT-4o and Claude 3.5, achieve relatively moderate to high ROUGE scores (0.4 to 0.5), especially in genres such as `news` and `biography`, suggesting a high degree of lexical overlap with the human reference and thus reliable SEE.

We also evaluate the lexical and semantic diversity of the five summaries per document (one human + four model-generated) using Self-BLEU (Zhu et al., 2018) and mean pairwise cosine similarity, in Figure 11 and Figure 12 respectively.

B.2 Instruction for Summary Crowdsourcing

Overall you will be assigned eight texts (from different genres) and asked to complete the following two tasks, which in total should take 2-3 hours. You will use only this form to record all the eight summaries you write.
1. Read the document
2.Write a one sentence summary for the document

Figure 8: Instruction for summary crowdsourcing.

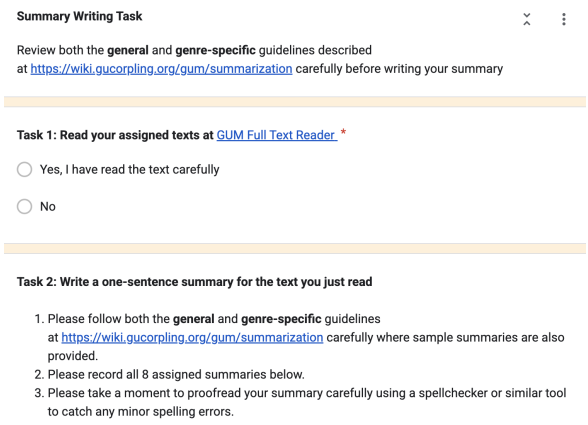


Figure 9: A screenshot of instructions given to the annotators.

Self-BLEU measures lexical diversity by averaging BLEU scores across all summary pairs. It ranges from 0 (completely diverse) to 1 (identical), with lower values indicating greater diversity. As indicated in Figure 11, Self-BLEU scores remain low (typically < 0.10) across genres, suggesting that the summaries are lexically diverse and not overly repetitive.

Cosine similarity, computed over sentence embeddings, captures how semantically similar the summaries are. It ranges from –1 to 1, with higher values indicating greater semantic alignment. In Figure 12, we observe high cosine similarity scores (e.g., 0.7-0.8) for most genres, reflecting strong consensus among human and model summaries in what content to include. In contrast, lower similarity scores for genres like `fiction` and `conversation` (e.g., 0.66–0.70) suggest a greater variability in what is considered salient in these genres. Together, these results confirm that the summaries used in our pipeline are both high-quality and sufficiently diverse to support robust salient entity extraction.

D Summary alignment interface

To inspect, evaluate and correct summary alignment, we used the GitDOX interface (Zhang and Zeldes, 2017), shown in Figure 13. The summaries are shown at the top of the interface and can be scrolled through. Coreferring mentions receive boxes with identical color borders, and are highlighted in yellow on hover ('Greek Court' in the example). Entities that appear in the currently selected summary have red text at their first mention, for example 'Zeus', while entities that are mentioned in a different summary but not the current one have blue text on their first mention. The total number of summaries that an entity is aligned to is shown with a plus and a number indicator, for example '+1' for Zeus, which appears only in summary1, but '+5' for the court, which appears in all five summaries.

E LLM Graded Entity Salience Prediction

We show the prompt for extracting and scoring salient entities in Figure 14. The prompt for predicting entity alignment is in Figure 15. We set a `temperature` of 0.2 to encourage deterministic outputs by focusing on the most probable responses. A `max_tokens` value of 300 limits verbosity and controls token usage, while a `top_p` of 0.7 reduces randomness, ensuring the model prioritizes relevant and reliable predictions.

F Additional Analyses on Graded Entity Salience Prediction

F.1 LLM Performance across Genres

Figure 16 shows that all LLMs perform relatively well on structured genres like `interview`, `news`, and `textbook`, which provide clear organizational cues (e.g. section titles and headlines), enabling easier salience scoring. In contrast, performance drops significantly in `conversation` and `vlog`, even for powerful models like GPT4o 3 shot, despite being provided with in-context examples. This disparity can be attributed to the unstructured and dynamic nature of these genres: conversations feature fragmented speech, rapid topic shifts, and implicit references, while vlogs often revolve around subjective, informal storytelling. The lack of clear salience signals in these genres makes it challenging for models to align their scores with human judgments.

The RMSE plot in Figure 17 reveals a contrasting trend, where unstructured genres like `conversation` and `vlog` perform well, with lower RMSE values compared to structured genres like `interview` or `textbook`. This suggests that while models struggle to rank entities accurately in these unstructured genres (as seen in Spearman correlations), they tend to predict scores that are numerically closer to the annotated values. This may occur because unstructured genres often feature a smaller range of salience variation, with fewer highly salient entities and many entities scored

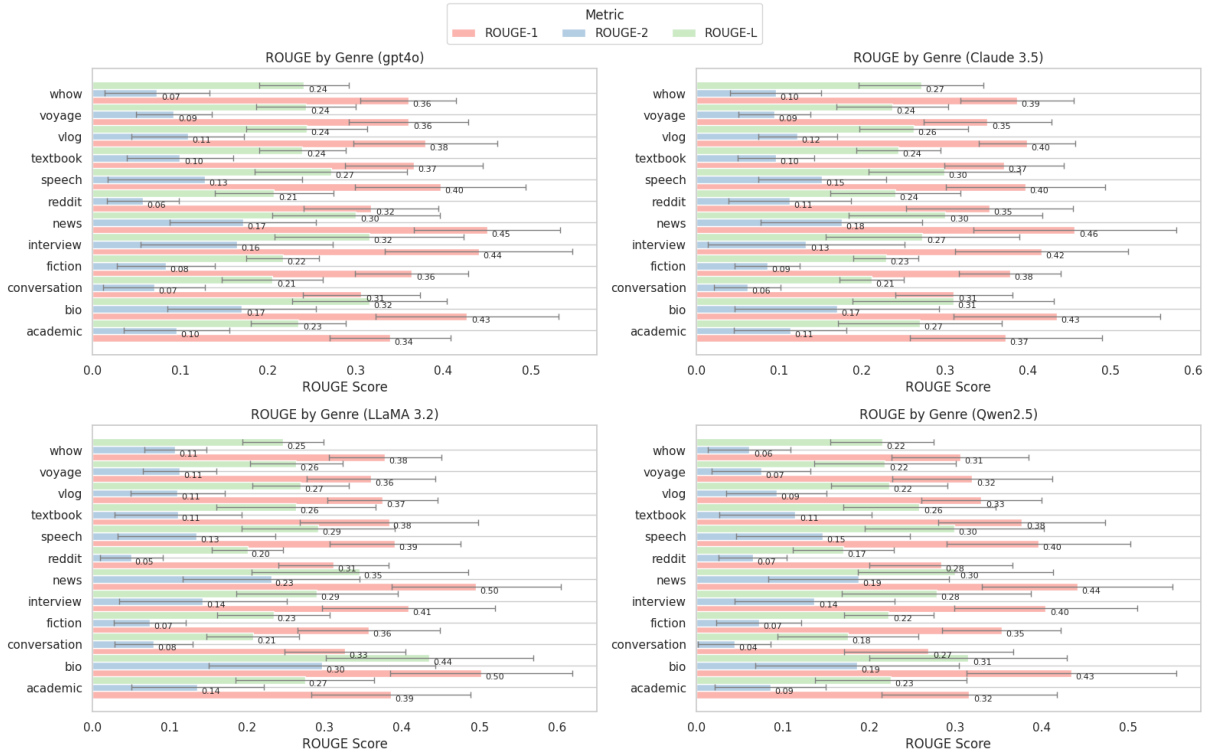


Figure 10: ROUGE-1, ROUGE-2, and ROUGE-L scores (with standard deviation) for summaries generated by four models (`GPT-4o`, `Claude 3.5`, `LLaMA 3.2`, and `Qwen2.5`), evaluated against human-written summaries across genres in the `training` set. Each subplot shows genre-level performance for one model.

similarly. Consequently, even when the ranking is incorrect, the predicted scores remain close to the annotated values, resulting in lower RMSE.

F.2 Confusion Matrices for all Models

The confusion matrices in Figure 18 reveal a general trend: the models in (a), (b), and (c), which correspond to GPT4o 3-shot, Llama 3-2, and Mistral 7B, tend to struggle with mid-range salience levels (scores 2 and 3), often misclassifying them as higher salience scores such as 4 or 5. In contrast, Stanza (d) and Ensemble (e) models demonstrate a stronger tendency to predict higher salience scores (4 or 5) more frequently. This is likely due to the training data for Stanza and Ensemble being skewed toward entities with higher salience scores, leading these models to overfit on these dominant categories. This observation underscores a potential limitation in the training data distribution, which might favor higher salience entities and influence the predictive bias of these models.

F.3 Model Performance across Entity Types

Analysis of model performance across entity types (Figure 19) reveals several consistent patterns. All models show stronger performance on concrete entities like ORGANIZATION and PERSON compared to abstract or temporal entities. While LLMs (GPT4o and Llama) exhibit higher recall than precision for PERSON entities, suggesting a tendency to overpredict these entity types, both Stanza and Ensemble demonstrate more balanced precision-recall trade-offs across entity types. Most models struggle with ABSTRACT entities, showing consistently lower F1 scores for this category compared to other entity types. This pattern suggests that identifying salient abstract concepts remains a key challenge across different architectural approaches.

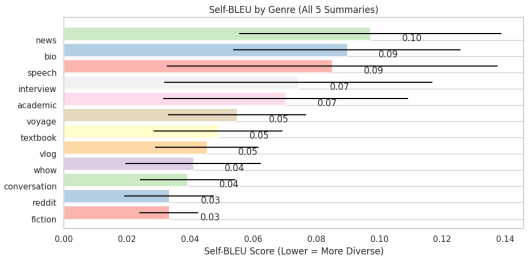


Figure 11: Self-BLEU scores (with standard deviation) for each genre, computed across all five summaries (1 human + 4 model-generated). Lower Self-BLEU values indicate greater lexical diversity among the summaries.

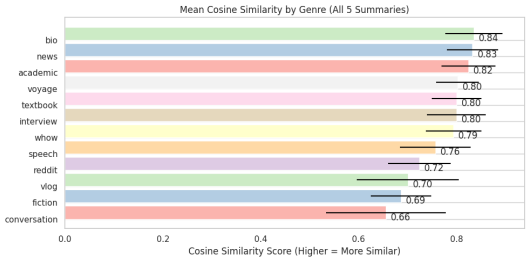


Figure 12: Mean pairwise cosine similarity (with standard deviation) between all five summaries per document, averaged by genre. Cosine similarity captures semantic similarity based on sentence embeddings, with values closer to 1 indicating stronger agreement in meaning.

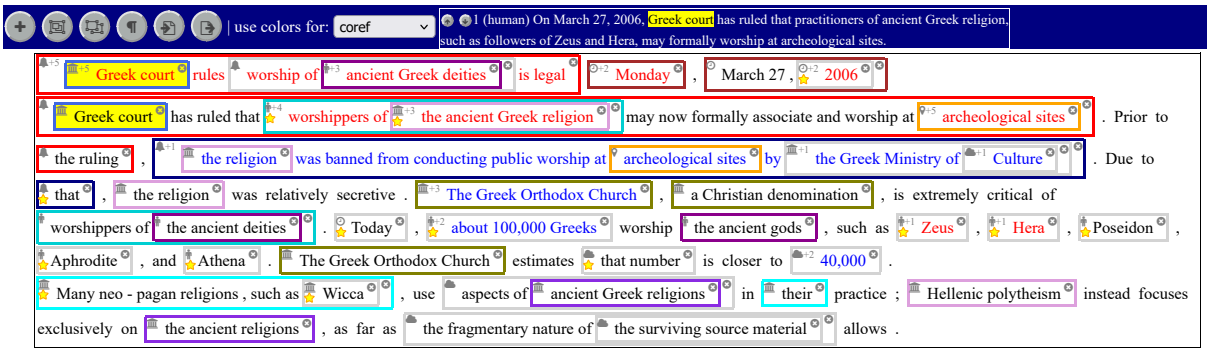


Figure 13: GitDOX annotation interface for summary entity alignment.

Prompt for Predicting Salient Entities

List salient entities in the following document, such as people, organizations, places, or other notable subjects. For each entity, provide a relevance score from 1 to 5, where 5 indicates the most relevant entities in the context of the document, and 1 indicates less relevant but still notable entities.
Document: {doc_text}

Here are 3 example documents along with their salient entities to help guide you:
Example Document 1: {example_doc_1}

Salient entities:
[('the division of labor', 4), ('production', 3), ...]

Figure 14: Prompt for predicting salient entities (3-shot).

Prompt for Predicting Entity Alignment

Consider the following document and summary:
Document: {doc_text}
Summary: {summary}
For each of the following entities in the document, if it matches with (refers to) the ones in the summary, return the exact same entity in the phrase in bullet points. Otherwise, don't return anything. When matching, please also consider synonyms or alternative phrases that refer to the same entity. If a speaker says 'I' or is mentioned as 'you', then the speaker's name or label is considered mentioned (e.g. Kim).
Entities: {entities}
Be very precise and only return entities that match those in the summary. Do not add extra or unrelated entities.

Figure 15: Prompt for predicting entity alignment.

G License and Copyright

All human-written summaries were collected with explicit consent and released under a CC0 license. Annotators were informed that their summaries would be publicly shared for research purposes and were given the option to remain anonymous or receive attribution. All annotators consented to data release. The consent form is shown in Figure 20.

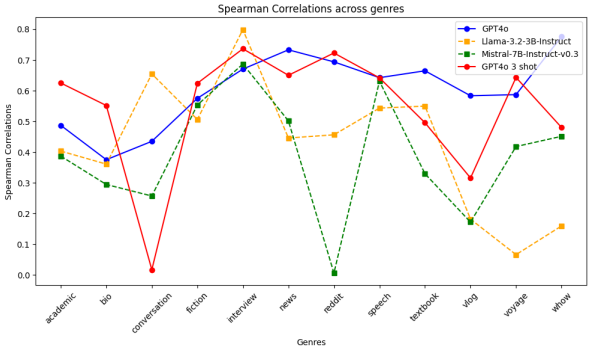


Figure 16: Spearman correlation with LLM scores across 12 genres.

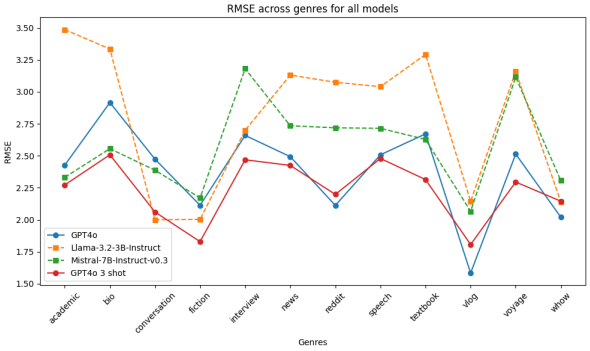
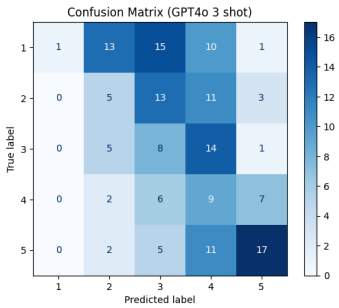
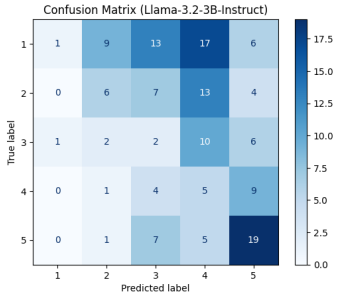


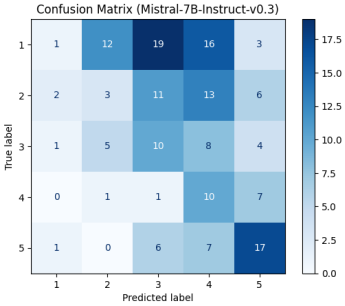
Figure 17: RMSE scores with LLM scores across 12 genres.



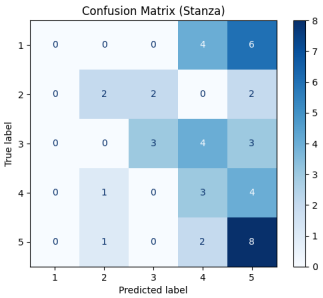
(a) GPT4o 3-shot



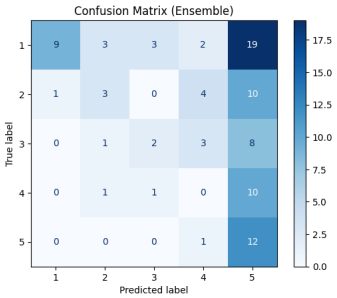
(b) LLama 3-2



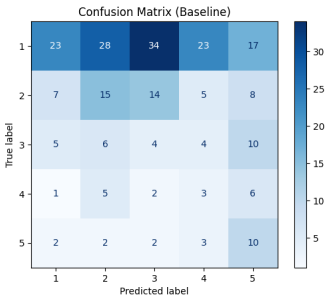
(c) Mistral 7B



(d) Stanza

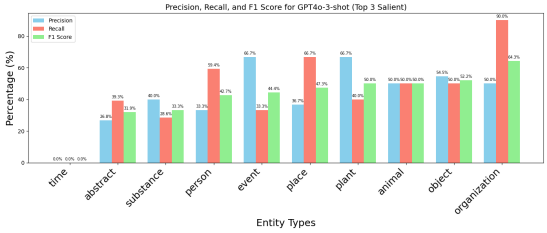


(e) Ensemble

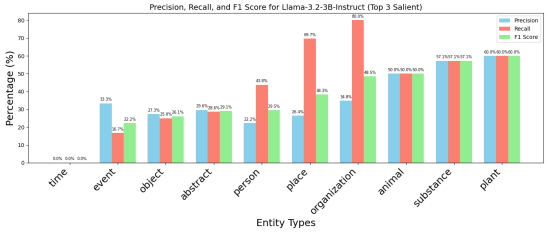


(f) Baseline

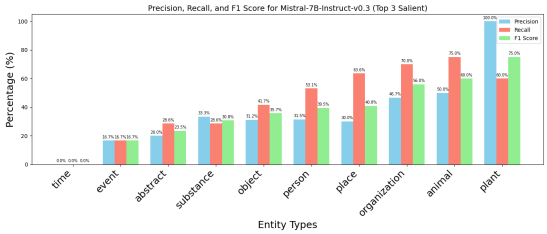
Figure 18: Confusion matrices for all models.



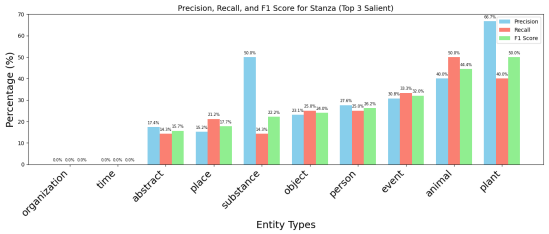
(a) GPT4o 3-shot



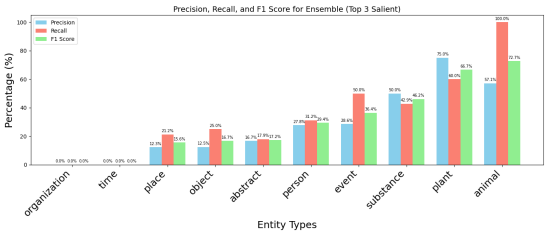
(b) Llama



(c) Mistral



(d) Stanza



(e) Ensemble

Figure 19: Performance of all models across entity types (Top 3 Salient Entities)

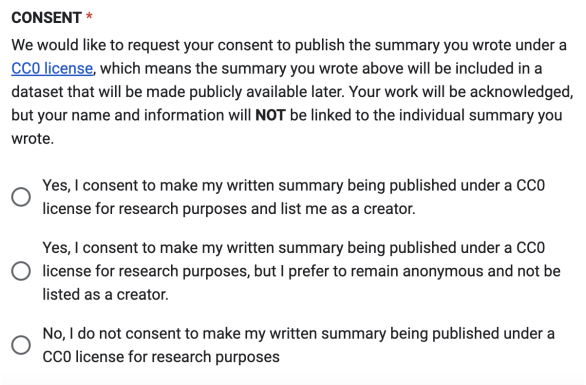


Figure 20: A screenshot of the consent form for the annotators.